

Egocentric Action-aware Inertial Localization in Point Clouds with Vision-Language Guidance

Mingfang Zhang¹, Ryo Yonetani², Yifei Huang^{1*}, Liangyang Ouyang¹, Ruicong Liu¹, Yoichi Sato¹

¹The University of Tokyo, ²CyberAgent AI Lab

{mfzhang,hyf,oyly,lruicong,ysato}@iis.u-tokyo.ac.jp, yonetani_ryo@cyberagent.co.jp

Abstract

This paper presents a novel inertial localization framework named *Egocentric Action-aware Inertial Localization (EAIL)*, which leverages egocentric action cues from head-mounted IMU signals to localize the target individual within a 3D point cloud. Human inertial localization is challenging due to IMU sensor noise that causes trajectory drift over time. The diversity of human actions further complicates IMU signal processing by introducing various motion patterns. Nevertheless, we observe that some actions captured by the head-mounted IMU correlate with spatial environmental structures (e.g., bending down to look inside an oven, washing dishes next to a sink), thereby serving as spatial anchors to compensate for the localization drift. The proposed EAIL framework learns such correlations via hierarchical multi-modal alignment with vision-language guidance. By assuming that the 3D point cloud of the environment is available, it contrastively learns modality encoders that align short-term egocentric action cues in IMU signals with local environmental features in the point cloud. The learning process is enhanced using concurrently collected vision and language signals to improve multimodal alignment. The learned encoders are then used in reasoning the IMU data and the point cloud over time and space to perform inertial localization. Interestingly, these encoders can further be utilized to recognize the corresponding sequence of actions as a by-product. Extensive experiments demonstrate the effectiveness of the proposed framework over state-of-the-art inertial localization and inertial action recognition baselines. Project page: <https://github.com/mf-zhang/Ego-Inertial-Localization>.

1. Introduction

Inertial localization has emerged as a pivotal technology for human tracking across a wide range of various applications, including personal navigation [22] and augmented reality [51]. Inertial measurement units (IMUs) are commonly embedded in wearable devices such as smartphones, smart glasses [12], and headsets [45]. When affixed to

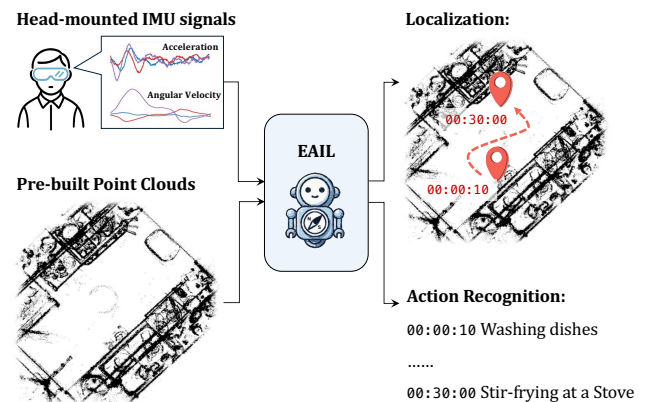


Figure 1. **Egocentric Action-aware Inertial Localization (EAIL)**. Our framework leverages egocentric action cues obtained from the head-mounted IMU to perform inertial localization in the environmental 3D point cloud. The corresponding sequence of actions can also be recognized as a by-product.

human bodies, IMUs can capture acceleration and angular velocity to record 3D human movements. Compared to vision-based localization methods [28, 39], inertial localization enables user tracking in an energy-efficient and privacy-preserving manner.

Despite the advantages, human inertial localization remains highly challenging primarily due to two reasons. The first one is *the trajectory drift caused by IMU noise*. IMU sensor noise causes small uncertainties in measurements that accumulate over time, which can ultimately lead to significant trajectory drift [58]. Conventional step detection methods [3, 64] struggle to generalize to the noise of irregular movements, while recent data-driven approaches [22, 37, 54, 66], which predict velocity from IMU signals, can suffer from accumulated estimation errors.

The other challenge lies in *the complexity of human actions*. Human body-mounted IMUs capture signals not only from displacement-related movements like walking and stopping, but also from action-induced motions that do not involve actual positional changes, such as head-swiveling during cooking and cleaning. These additional

*Corresponding author.

motion signals can complicate IMU signal processing and make inertial localization further difficult. Existing datasets [5, 22, 62] and approaches [24, 62] focus mostly on human walking scenarios, limiting the ability of state-of-the-art inertial localization methods to handle real-world human action variability.

Nevertheless, we argue that human actions can rather act as a salient locational cue to mitigate the trajectory drift challenge if properly taken into account. For example, washing dishes often takes place near a sink, while frying often occurs near a stove. Our key insight is that extracting such actions from IMU data can help to deduce the spatial regions where they are likely to occur.

In this work, we present a novel framework named *Ego-centric Action-aware Inertial Localization (EAIL; see also Fig. 1)*. While assuming that the 3D point cloud of the environment is readily available using off-the-shelf 3D scanners, this framework leverages egocentric action cues extracted using head-mounted IMUs to localize the target individual in the environmental point cloud.

The key technical contribution is a hierarchical modality alignment technique that effectively extracts the correlation between the action cues and the point cloud with vision-language guidance. Specifically, we first learn the encoders for short-term IMU signals and local point clouds to be aligned with the corresponding egocentric point-of-view images and textual descriptions of observed actions in a contrastive fashion. By incorporating guidance from pretrained vision-language models, we achieve improved multimodal feature alignment. The learned IMU and point-cloud encoders are then invoked in spatio-temporal reasoning modules to jointly predict possible actions occurring at each moment and the locations in the global point cloud where the predicted actions can be observed. While our primary goal is the localization in the latter part, the predicted locations can also improve inertial action recognition as a beneficial by-product.

Extensive evaluations on the EgoExo4D dataset [18] validate that our framework achieves state-of-the-art performance in both inertial localization and inertial action recognition compared to [24, 41, 66, 69]. In summary, our main contributions are as follows:

- We introduce EAIL, a novel inertial localization framework that leverages egocentric action cues from head-mounted IMU signals to localize target individuals within a 3D point cloud.
- We develop a hierarchical modality alignment technique that learns the correlation between the inertial egocentric action cues and the environmental point cloud to perform inertial localization as our primary goal, and action recognition as a by-product.
- Extensive evaluations demonstrate that our framework achieves state-of-the-art results in both inertial localiza-

tion and inertial action recognition in diverse settings.

2. Related Work

2.1. Human Inertial Localization

Human inertial localization has been an active research area. Leveraging both visual and inertial data, accurate localization can be achieved by perceiving and modeling the surrounding environment [31, 42, 49]. Visual-inertial fusion can also simultaneously perform localization and body poses estimation [35, 63]. However, visual data availability is not guaranteed at all times, especially in human-centric applications where privacy concerns are essential.

Traditional inertial localization methods like Pedestrian Dead Reckoning (PDR) [3, 53, 64] estimate trajectories by detecting steps and estimating step length and heading. While effective in structured environments, these approaches struggle with noise from irregular human motion. Data-driven inertial navigation methods [4, 22, 66] mitigate sensor noise by learning to estimate velocity from IMU signals. However, when they cumulate the estimated velocity to predict human locations, the estimation errors can lead to significant trajectory drift.

To mitigate the drift problem, external signals such as GPS [29], Wi-Fi [23], Bluetooth [32], user history [65], and activities [13, 14] have been used to provide anchor points for human localization. However, these solutions introduce dependencies on specific infrastructures or hand-defined landmarks. More recently, NILoc [24] takes a different approach by directly predicting user locations without explicit velocity integration, but it requires scene-specific training and lacks adaptability across diverse environments.

Finally, beyond these methodological challenges, existing datasets [5, 22] and approaches [24, 62] are primarily centered around displacement-centric activities such as walking and stopping. This focus overlooks a broader spectrum of complex human actions, thus limiting the models' abilities to manage real-world human motion variability. In contrast, our framework addresses the trajectory drift problem by leveraging egocentric action cues from IMU signals, to localize the user within a point cloud.

2.2. Egocentric Multimodal Alignment

Recent advancements in head-mounted devices [12, 51] have generated significant interest in understanding human activities from an egocentric perspective. Several large-scale egocentric datasets [8, 17, 18, 25] have been introduced, offering rich multimodal data including video, gaze, audio, language, and IMU signals. These datasets provide a foundation for developing multimodal human understanding methods [16, 34, 36, 41].

Aligning egocentric visual data with textual action descriptions enhances comprehension and broadens applica-

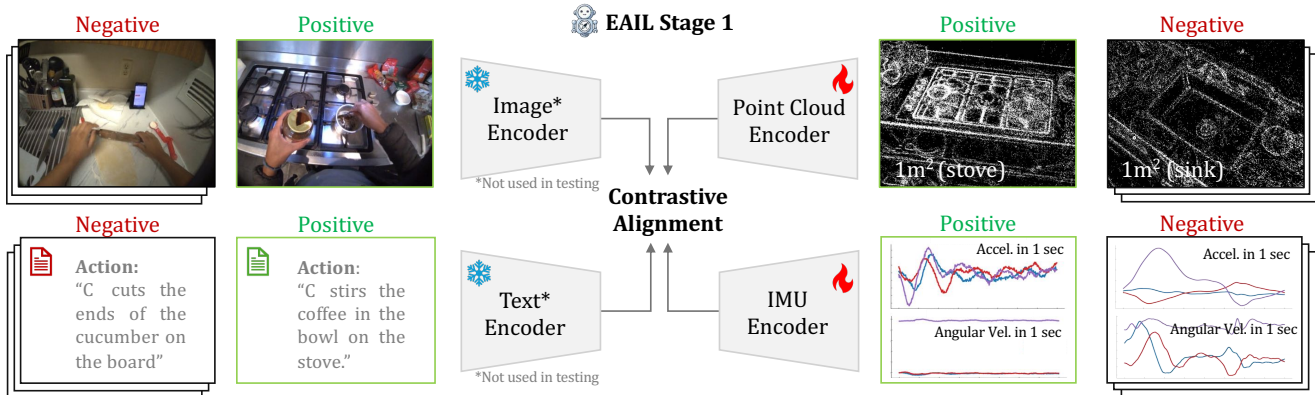


Figure 2. **Short-Term Action-Location Alignment.** In this first stage, our objective is to train a point cloud encoder and an IMU encoder using contrastive learning. A positive sample set consists of simultaneous multimodal data samples at the human’s location. Negative samples are generated by randomly sampling data from other times and locations.

tion possibilities. Large-scale video-language pretraining, specifically tailored to egocentric data, has been pioneered by works like EgoVLP [33, 47] and HierVL [2]. In addition to the human language narrations, LaViLa [70] refines these narrations with Large Language Models (LLMs). Recap [27] and Vinci [26] align fused languages for long videos. EMBED [11] and EgoInstructor [60] use rules or retrieval models to add additional training data. Subsequent works further incorporate additional egocentric-specific cues such as hand [67] and audio [6]. These approaches have demonstrated that egocentric-specific pre-training yields transferable representations for various downstream tasks, such as question answering [15], robot manipulation [46], and narration-based segmentation [52].

Aligning vision and language embeddings with 3D environments [38, 61, 68] makes open-vocabulary affordance detection possible. Affordance denotes potential actions in the environment that an agent can perform to an object or in an area. [9] provides accurate interaction annotations for real-world 3D indoor scenes. Works like [19, 40, 55] utilize free-form queries describing geometry and affordances, leveraging the pretrained CLIP [50] model, to retrieve 3D instances. These studies show that a strong correlation between actions and environments can be learned through multimodal alignment.

Inspired by these advances, our framework introduces learning the multimodal alignment between the human action cues and the environmental structures, to serve as natural anchors in the scene to predict human trajectories.

3. Problem Setting

We address the task of inertial localization using head-mounted devices equipped with IMUs. We assume that the 3D point cloud of the environment is available using consumer-grade 3D scanner apps [1, 30]. Our primary goal is to predict the sequence of the device user’s locations in

the point cloud.

Formally, let $\mathcal{M} = [\mathbf{M}_1, \dots, \mathbf{M}_T]$ represent the sequence of IMU data over T seconds, where each \mathbf{M}_t corresponds to the IMU signals captured within a single second. These signals comprise a sequence of 6-DoF sensor readings, specifically consisting of 3-axis accelerations and 3-axis angular velocities. We also denote the point cloud of the environment by \mathcal{P} . The sequence of locations of the target individual on the ground, $\mathcal{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_T]$, $\mathbf{z}_t \in \mathbb{R}^2$, is represented in the same coordinate system as that of \mathcal{P} . Our inertial localization method will learn a mapping $(\mathcal{M}, \mathcal{P}) \mapsto \mathcal{Z}$.

Note that this problem setup is different from existing inertial navigation (*e.g.*, [22]) and inertial localization [24]. Inertial navigation methods just predict velocity and thus require a known initial position to perform localization, making them prone to cumulative drift. Existing inertial localization [24] operates in a scene-specific manner, training and evaluating models separately for each environment while disregarding explicit environmental structure. In contrast, our approach incorporates the 3D point cloud \mathcal{P} , enabling localization without requiring environment-specific training.

4. The EAIL Framework

The proposed EAIL framework leverages action cues from head-mounted IMU signals to localize target individuals within the 3D point cloud. It proceeds in two stages:

- **Stage 1: Short-term action-location alignment** (Fig. 2). We first train an IMU encoder for processing short-term IMU signals (*e.g.*, spanning 1 second) as well as a point-cloud encoder for local point clouds (*e.g.*, 1 m²). Here we propose a novel modality alignment technique that aligns the features of those IMU and point cloud inputs with the corresponding egocentric point-of-view images and text annotations of the actions taken by the target individual,

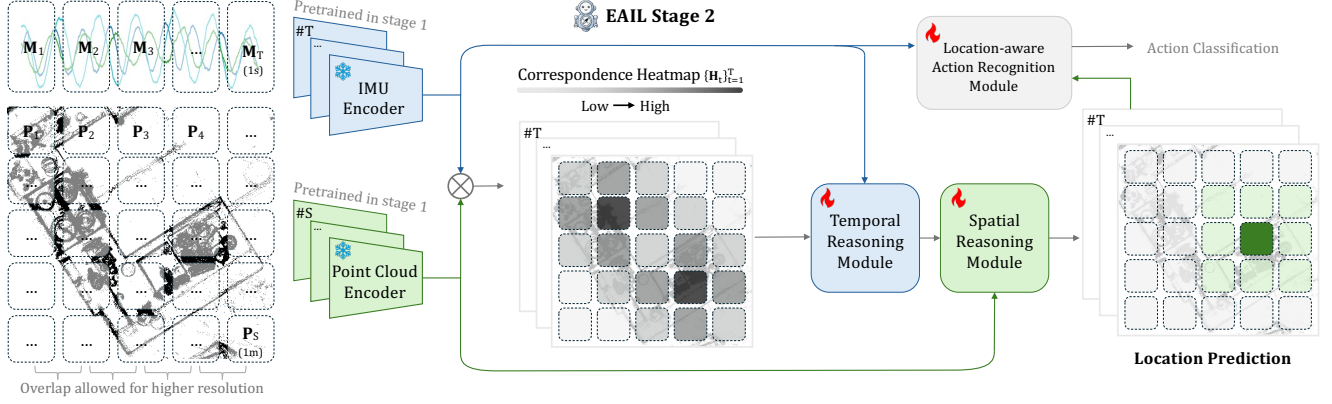


Figure 3. **Sequential Motion Localization.** In this second stage, we generate a sequence of the user’s locations and actions over T seconds using a series of IMU signals ($\{\mathbf{M}_t\}_{t=1}^T$) alongside the point cloud of the entire scene. This point cloud is divided into S local segments ($\{\mathbf{P}_s\}_{s=1}^S$). The IMU encoder and the point cloud encoder are frozen networks pre-trained in Stage 1 for efficient spatial-temporal reasoning in large-scale point clouds.

which makes the trained encoders ‘action-aware’. Note that these egocentric images and text annotations are necessary only in this training stage.

- **Stage 2: Sequential motion localization** (Fig. 3). We then learn spatiotemporal reasoning modules that invoke the IMU/point-cloud encoders to predict the sequence of target individual’s locations (and the sequence of their actions as a bonus; see Sec. 4.2).

4.1. Stage 1: Short-term Action-Location Alignment

In the first stage, we aim to train a short-term IMU signal encoder and a local point cloud encoder. In the training, we focus on capturing action semantics to establish a meaningful correspondence between the two modalities. These correlations, such as the association between dish-washing motion and the structure of a sink, can serve as natural anchors to enhance downstream localization accuracy.

To achieve this goal, we employ multimodal contrastive learning. Egocentric datasets [8, 17, 18] usually include vision and text data alongside IMU data. We conduct contrastive alignment across four modalities: images, textual action descriptions, IMU signals, and point clouds, as shown in Fig. 2. In this way, we can utilize the robust capabilities of pre-trained and pre-aligned vision-language models [7, 43, 50]. The action semantics embedded in vision and text modalities aid to guide the training of the IMU and the point cloud encoders. Note that the image and the language modalities are not required during testing.

Specifically, for each 1-second time segment, we extract a synchronized set of inputs: an ego-view image frame \mathbf{I}_t , an action caption \mathbf{L}_t , an IMU segment \mathbf{M}_t , and a local $1m^2$ region of point cloud \mathbf{P}_t . We utilize off-the-shelf pre-trained vision-language encoders E_I and E_L to encode the image and the text data, while we train the IMU and the point cloud

encoders, E_M and E_P . Then, we apply a pairwise contrastive loss L_c across the generated features \mathbf{F}^I , \mathbf{F}^L , \mathbf{F}^M , and \mathbf{F}^P derived from the four modalities:

$$L_{stage1} = \alpha L_c(\mathbf{F}^I, \mathbf{F}^M) + \beta L_c(\mathbf{F}^I, \mathbf{F}^P) + \theta L_c(\mathbf{F}^L, \mathbf{F}^M) + \delta L_c(\mathbf{F}^L, \mathbf{F}^P) + \gamma L_c(\mathbf{F}^M, \mathbf{F}^P). \quad (1)$$

Through this process, our IMU and point cloud encoders are trained to generate features in a shared embedding space where multimodal features are connected through human action semantics. This enables the retrieval of local point cloud segments with short-term IMU signals. It can provide a preliminary estimate of the user’s location with IMU signals. However, given that similar motion patterns might occur in multiple locations, we introduce the second stage of our framework to address these ambiguities.

4.2. Stage 2: Sequential Motion Localization

In Stage 2, we build on the pre-trained short-term IMU encoder and the local point cloud encoder, and invoke them for processing sequential motion captured over T seconds of IMU signals \mathcal{M} and the global point cloud \mathcal{P} of the entire scene to predict the user’s trajectory \mathcal{Z} .

4.2.1. Spatiotemporal reasoning for trajectory prediction

Predicting user trajectories within large-scale point clouds requires fine-grained temporal and spatial reasoning. To achieve this, as shown in Fig. 3, we decompose the T -second IMU signals \mathcal{M} into short sub-sequences, $\{\mathbf{M}_t\}_{t=1}^T$. Concurrently, the global point cloud is uniformly partitioned into S local segments, denoted as $\{\mathbf{P}_s\}_{s=1}^S$. Then, we apply the frozen E_M to extract IMU features $\{\mathbf{F}_t^M\}_{t=1}^T$. In parallel, the frozen E_P processes local point cloud segments to produce point cloud features $\{\mathbf{F}_s^P\}_{s=1}^S$.

Given that IMU features \mathbf{F}^M and point cloud features \mathbf{F}^P are aligned due to the Stage 1 contrastive learning process, we can generate a sequence of *correspondence heatmaps* $\{\mathbf{H}_t\}_{t=1}^T$ by calculating the similarity between these two feature sets. Regions with a high score on these heatmaps indicate a high likelihood that human motion occurs in these spatial regions. For instance, “washing” would align closely with a “sink” region, while “walking” covers “open areas”.

Building on this, we design a Temporal Reasoning Module and a Spatial Reasoning Module. The temporal reasoning module takes the correspondence heatmaps $\{\mathbf{H}_t\}_{t=1}^T$ and IMU features $\{\mathbf{F}_t^M\}_{t=1}^T$ as inputs and employs a 3D convolutional network to reason across the temporal dimension, generating refined features \mathbf{F}^R . Next, \mathbf{F}^R is inputted, along with the point cloud features $\{\mathbf{F}_s^P\}_{s=1}^S$, into the spatial reasoning module. This module uses another dilated 3D convolutional network to reason in the 2D spatial dimension. The final output is the predicted user trajectory \mathcal{Z} .

For the trajectory \mathcal{Z} , we frame the prediction as an S -class classification problem. The goal is to find a sequence of point cloud segments that are nearest to the actual user trajectory. We use cross-entropy loss for training:

$$L_{traj} = - \sum_{t=1}^T \sum_{s=1}^S \mathbf{y}_{t,s} \log(\hat{\mathbf{y}}_{t,s}), \quad (2)$$

where $\mathbf{y}_{t,s}$ is the true label indicating the user’s nearest point cloud segment and $\hat{\mathbf{y}}_{t,s}$ is the predicted probability for time step t and segment P_s .

4.2.2. Location-aware action recognition

In the Location-aware Action Recognition Module, we leverage the predicted user locations to enhance inertial action recognition. Specifically, for each time step t , we take all the predicted location probabilities $\hat{\mathbf{y}}_{t,s}$ in a scene, collectively represented by heatmaps $\{\mathbf{H}'_t\}_{t=1}^T$, and use them as a spatial attention over the point cloud features $\{\mathbf{F}_s^P\}_{s=1}^S$. We then blend these spatial features with IMU features $\{\mathbf{F}_t^M\}_{t=1}^T$ through addition. Finally, a multi-layer perceptron maps the fused representation to action likelihood. The training is supervised by a cross-entropy loss:

$$L_{action} = - \sum_{t=1}^T \sum_{c \in \mathcal{C}} \mathbf{y}_{t,c} \log(\hat{\mathbf{y}}_{t,c}), \quad (3)$$

where $\mathbf{y}_{t,c}$ represents the true label for the action class c at time t , and $\hat{\mathbf{y}}_{t,c}$ is the predicted probability.

In summary, our whole model in Stage 2 is supervised by $L_{stage2} = L_{traj} + L_{action}$.

5. Experiments

5.1. Experimental Setup

Dataset We use the EgoExo4D dataset [18] that provides synchronized egocentric video and IMU signals recorded

with Aria glasses [12]. Prior to data collection, environment point clouds were built, and action captions were annotated afterward. For this paper, we used the cooking activities subset, which includes 173 participants across 60 kitchens, totaling 564.13 hours of recordings. The activity area for these cooking activities averaged around 2.8 meters per side, with the largest spanning 6.15 meters. Each local 1 m^2 point cloud is sub-sampled to contain 8192 points, and the IMU signals are recorded at an 800 Hz sample rate. The dataset is labeled with 35 distinct action classes for classification tasks. We divided the data into training, test-seen, and test-unseen sets. The test-unseen set consists exclusively of environments and participants not present in the training set.

Implementation Details In Stage 1, we employed an enhanced version of the CLIP model [43] as the vision-language encoder. Specifically, we used ViT-Base [10] as the image encoder and the Base Text Transformer [59] from CLIP for text encoding. For the point cloud encoder, we adopted PointNet++ [48], and the IMU encoder was built upon ResNet18-1D [20]. The IMU signals are pre-processed following [22] and downsampled to 400 Hz. In Stage 2, our model operated with a sequence length $T = 10$, uniformly partitioning the point cloud into $S = 20 \times 20 = 400$ local segments. We trained Stage 1 for 250 epochs and Stage 2 for 100 epochs, using a batch size of 64, a learning rate of 10^{-3} , and the AdamW optimizer. The loss parameters were set as follows: $\alpha = 0.1$; $\beta, \theta, \delta, \gamma = 1$.

Evaluation Metrics For the localization task, we report the success rate (%) at error distance thresholds of 0.2 m, 0.4 m, and 0.6 m following NILoc [24]. Additionally, we include the Relative Score (RS) metric, which measures the proportion of heatmap positions with lower confidence than the ground truth location. This metric indicates the relative ranking of the prediction at the target point. For action classification, we evaluate performance using top-1 and top-5 accuracy metrics. These scores are assessed under two setups: “seen rooms” where the localization is performed in the environments present in the training dataset and “unseen rooms” where environments are otherwise new.

5.2. Inertial Localization Results

Baselines RoNIN [22] learns to predict velocity from IMU signals. We select ResNet-50 as the most effective backbone. IMUNet [66] introduces a new architecture and implements it with various backbones [56, 57]. NILoc [24] directly predicts user location with IMU signals. We choose to use ResNet-18 to encode IMU signals to maintain consistency with our framework. We train all the aforementioned methods on the EgoExo4D dataset.

Table 1. **Inertial Localization Results.** We evaluate the accuracy using two metrics: the localization success rate (%) at various error distance thresholds and the Relative Score (RS) metric for localization likelihood prediction (methods that do not generate likelihood predictions are not evaluated with RS). For both metrics, higher values indicate better performance.

| Method | Seen Rooms | | | | Unseen Rooms | | | |
|------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | 0.2m | 0.4m | 0.6m | RS | 0.2m | 0.4m | 0.6m | RS |
| MnasNet[57] | 2.95 | 7.57 | 12.75 | / | 2.16 | 5.45 | 9.86 | / |
| EfficientNet[56] | 3.10 | 7.77 | 12.71 | / | 2.54 | 6.68 | 10.84 | / |
| IMUNet[66] | 3.74 | 10.15 | 17.23 | / | 3.40 | 9.17 | 14.63 | / |
| RoNIN[22] | 4.86 | 12.77 | 20.65 | / | 3.96 | 9.52 | 15.65 | / |
| NILoc[24]+ | 17.03 | 41.31 | 74.15 | 88.17 | 13.32 | 37.85 | 69.21 | 84.08 |
| Ours | 43.86 | 70.15 | 89.60 | 96.01 | 26.86 | 65.97 | 90.79 | 89.55 |

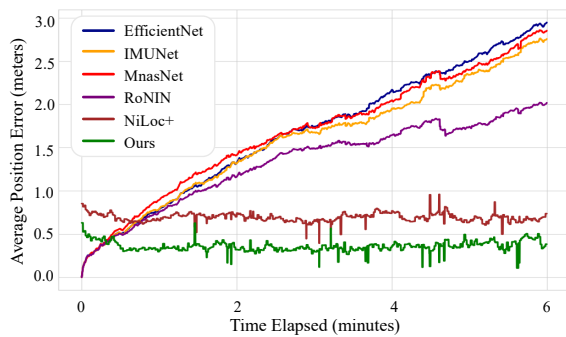


Figure 4. **Inertial Localization Error Over Time Elapsed.** While velocity accumulation-based methods experience significant trajectory drift, our approach remains accurate over time.

Velocity Cumulation versus Direct Location Prediction

Previous approaches can be divided into two categories. The first category starts with a given initial localization, estimates velocity using IMU signals, and accumulates the velocity to obtain the target’s location, including [22, 56, 57, 66]. The second category directly predicts the target’s location, including [24] and ours. As shown in Tab. 1, direct location prediction methods outperform velocity accumulation methods. This is because each recording in the dataset is relatively long, with an average duration of 525 seconds and a maximum of 2,526 seconds. Velocity accumulation methods are significantly affected by error accumulation over such long durations.

Trajectory Drift Fig. 4 illustrates the difference between the two categories of methods with the change in average position error over elapsed time. In this figure, velocity accumulation methods require a given initial position, resulting in zero error at time zero. However, due to error accumulation over time, their error increases rapidly, surpassing that of our method at around 30 seconds. In contrast, direct localization prediction methods do not require an initial position and maintain a stable error over time.

Table 2. **Inertial Action Recognition Results.** We evaluate performance using top1 and top5 accuracy metrics. Higher values indicate better performance.

| Method | Seen Rooms | | Unseen Rooms | |
|------------------|--------------|--------------|--------------|--------------|
| | top1 | top5 | top1 | top5 |
| DeepConvLSTM[44] | 15.20 | 43.27 | 12.47 | 36.86 |
| EVIMAE[69]– | 18.34 | 46.12 | 9.30 | 31.10 |
| IMU2CLIP[41] | 18.96 | 50.43 | 12.27 | 37.04 |
| Ours | 21.48 | 53.62 | 15.03 | 43.34 |

Direct Localization in Point Clouds NILoc [24] and our method both directly predict a user’s location from IMU signals. However, NILoc relies solely on inertial data, so it can only fit a single scene at a time. In their original work, they trained three separate models, each for a different scene. For a fair comparison, we developed NILoc+, which uses the same multi-scene training data as our Stage 2. Nevertheless, its lack of spatial awareness still leads to reduced accuracy, whereas our approach leverages point cloud structures to deliver robust inertial localization across diverse environments.

5.3. Inertial Action Recognition Results

Baselines DeepConvLSTM [44] uses convolutional networks and LSTMs to classify actions from IMU signals. EVIMAE [69] leverages the MAE framework [21] to learn representations from both video and multiple IMU signals for action classification. For comparison with our approach, we train a modified version of EVIMAE using data from a single IMU device, employing ViT-Base as the backbone. IMU2CLIP [41] uses a strategy similar to our Stage 1, employing a pretrained CLIP model [43, 50] to guide IMU feature extraction and fine-tuning with an MLP head for action recognition. For a fair comparison, we use the same backbone for each modality and the same training data as our method when training IMU2CLIP.

Table 3. **Ablation Studies.** We report the inertial localization accuracy and the inertial action recognition (\mathcal{A}) accuracy simultaneously.

| Method | Seen Rooms | | | | | | Unseen Rooms | | | | | |
|--|--------------|--------------|--------------|--------------|---------------------|---------------------|--------------|--------------|--------------|--------------|---------------------|---------------------|
| | 0.2m | 0.4m | 0.6m | RS | \mathcal{A} -top1 | \mathcal{A} -top5 | 0.2m | 0.4m | 0.6m | RS | \mathcal{A} -top1 | \mathcal{A} -top5 |
| <i>Modalities Engagement (visual and textual action description benefits localization)</i> | | | | | | | | | | | | |
| w/o vision language | 39.75 | 66.30 | 87.56 | 95.70 | 13.21 | 51.02 | 20.41 | 61.95 | 89.83 | 88.13 | 9.53 | 29.09 |
| w/ vision language | 43.86 | 70.15 | 89.60 | 96.01 | 21.48 | 53.62 | 26.86 | 65.97 | 90.79 | 89.55 | 15.03 | 43.34 |
| <i>Action Classification Task (explicit action supervision benefits localization)</i> | | | | | | | | | | | | |
| w/o action loss | 41.92 | 68.54 | 87.75 | 95.51 | / | / | 25.37 | 63.79 | 89.28 | 89.02 | / | / |
| w/ action loss | 43.86 | 70.15 | 89.60 | 96.01 | 21.48 | 53.62 | 26.86 | 65.97 | 90.79 | 89.55 | 15.03 | 43.34 |
| <i>Spatial and Temporal Reasoning</i> | | | | | | | | | | | | |
| w/o spatial | 38.68 | 66.13 | 87.78 | 95.05 | 21.20 | 52.96 | 25.03 | 59.39 | 83.54 | 88.56 | 14.40 | 43.77 |
| w/o temporal | 41.44 | 68.48 | 87.96 | 95.78 | 21.22 | 52.48 | 26.41 | 64.15 | 89.13 | 89.41 | 14.61 | 43.06 |
| w/ both | 43.86 | 70.15 | 89.60 | 96.01 | 21.48 | 53.62 | 26.86 | 65.97 | 90.79 | 89.55 | 15.03 | 43.34 |

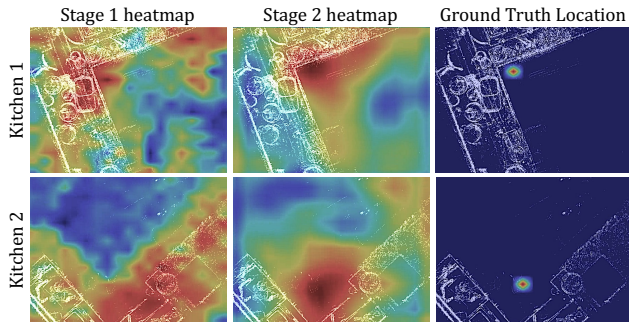


Figure 5. **Visualization of Heatmaps in Each Stage.**

Quantitative Evaluation As shown in Tab. 2, DeepCon-vLSTM performs relatively poorly due to its architecture’s difficulty in capturing complex action patterns effectively. Although EVIMAE gains from the advanced reasoning capabilities inherent in transformer networks, it still faces challenges when generalizing to unseen rooms. Much like our method, IMU2CLIP leverages the learning of action semantics from vision-language modalities; however, it doesn’t quite match the effectiveness of our approach. This is largely because our method incorporates spatial reasoning, utilizing environmental information to enhance action recognition performance.

5.4. Ablation Studies

Modalities Engagement for Action-aware Alignment

In Stage 1 of our framework, we focus on effectively training the IMU and the point cloud encoders to yield features that are aligned through action semantics. To accomplish this, we leverage the power of robust pre-trained and pre-aligned vision-language models, such as [43, 50]. These models excel at capturing intricate associations between visual content and corresponding textual descriptions. Our experiments, as detailed in Tab. 3, affirm the efficacy of in-

Table 4. **Location-Aware Action Recognition Ablation Study.** “PC” denotes point cloud features, and “LA” represents location attention.

| Method | Seen Rooms | | Unseen Rooms | |
|----------|--------------|--------------|--------------|--------------|
| | top1 | top5 | top1 | top5 |
| w/o PC | 18.81 | 50.23 | 12.37 | 38.21 |
| w/o LA | 17.88 | 46.22 | 10.43 | 35.63 |
| w/ LA PC | 21.48 | 53.62 | 15.03 | 43.34 |

corporating vision and language modalities. By doing so, our model is able to learn and align features across multiple modalities, significantly enhancing both localization and action recognition accuracy. Importantly, it’s worth noting that while the integration of vision and language modalities provides substantial benefits during the training phase, they are not required during the inference phase. Furthermore, even in scenarios where action caption annotations are unavailable in the training set, our method does not fail, ensuring reasonable accuracy without relying on complete annotation sets. This demonstrates the broad applicability and flexibility of our approach in real-world environments.

Action Classification Supervision in Stage 2

We incorporate action classification supervision in Stage 2. The results in Tab. 3 show that including the action loss leads to more accurate location predictions. We believe that explicitly learning action categories helps the model align human motion more effectively with the surrounding environment.

Spatial and Temporal Reasoning in Stage 2

In the second stage of our framework, we leverage a temporal reasoning module for comprehending a sequence of IMU signals, and a spatial reasoning module for understanding the global environmental point cloud. This dual-layered rea-

soning framework is essential, as evidenced in Tab. 3. It shows that when these two modules work in tandem, they align sequential motion data with the environmental settings. This holistic reasoning capability ensures the model generates a more precise and coherent trajectory prediction of the user’s movements within their environment.

Spatial Attention Benefits Inertial Action Recognition

After obtaining the inertial localization results, we find that incorporating surrounding environmental structure cues is beneficial for inertial action recognition. As shown in Tab. 4, using only IMU signals, we achieve results comparable to IMU2CLIP [41]. Naively incorporating global point cloud features with IMU features leads to a performance drop. In contrast, integrating point cloud features with predicted location attention with IMU features provides a clear performance improvement.

More Ablation Results in Supplementary Material

Further ablation results can be found in Tab. 5 in our Supplementary Material. These include analyses on different vision-language encoders in Stage 1, the preliminary location retrieval accuracy in Stage 1, different architecture designs in Stage 2, and different choices of sequence temporal length in Stage 2.

5.5. Qualitative evaluations

Visualization of Heatmaps in Each Stage Our framework is capable of producing highly interpretable intermediate results in the form of heatmaps. The heatmap from Stage 1 reflects the direct similarity strength between the features generated by the IMU encoder and the point cloud encoder. The heatmap from Stage 2 represents the localization likelihood at each location, considering both the motion sequence and the global scene structure. As shown in Fig. 5, in Kitchen 1, two peak points appear in the Stage 1 heatmap, indicating that the motion in the IMU signals could plausibly occur at either of these locations. However, after the spatiotemporal reasoning in Stage 2, our framework successfully identifies a single, distinct peak.

Comparison with Previous Methods In Fig. 6, we present a qualitative comparison with RoNIN[22] and NILoc[24]+. The recorded sequence features a person taking 12 minutes to make a cup of coffee. From the visualization results, we observe that RoNIN suffers from cumulative errors, causing its predicted location to drift outside the point cloud boundary after just 2 minutes. NILoc+, on the other hand, lacks spatial awareness and may produce predictions in physically implausible locations, such as areas already occupied by objects. In contrast, our method leverages action cues such as the upward head motion when fetching items from a cabinet, to establish action-environment correlations, enabling accurate localization.

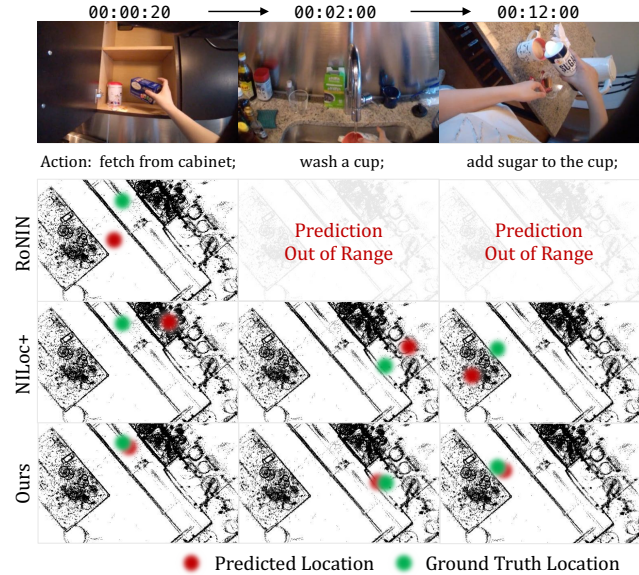


Figure 6. Qualitative Comparison with Previous Methods.

6. Limitations and Future Directions

While our method can robustly exploit head-mounted IMU signals for human localization within pre-built point clouds, it does hinge on several factors that present avenues for future research. First, the requirement of a complete and up-to-date 3D scan limits applicability in environments subject to frequent layout changes, motivating efforts on incremental or online map updates. Second, our reliance on action-environment correlations makes localization challenging when users remain idle (*e.g.*, standing still for extended periods). Integrating standard inertial navigation solutions and Kalman filters could help bridge gaps during low-motion segments. Finally, our experiments are based on IMU data from head-mounted devices, and substantially different sensor placements (*e.g.*, ankle or wrist) may necessitate model adaptations for robust performance.

7. Conclusion

We propose EAIL, a novel framework for inertial localization that uses egocentric action cues from head-mounted IMU signals to improve positioning in 3D point clouds. By aligning short-term IMU data with local structures and incorporating temporal-spatial reasoning, EAIL addresses challenges like trajectory drift and complex actions. Evaluated on the EgoExo4D dataset, it shows strong performance in trajectory prediction and action recognition. This highlights the potential of action-environment correlations as anchors for inertial localization, paving the way for advanced sensing and positioning methods.

8. Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP24K02956, JST ASPIRE Grant Number JPM-JAP2303, JST SPRING Grant Number JPMJSP2108.

References

- [1] Scaniverse. Accessed on March 7th, 2025. 3
- [2] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23066–23078, 2023. 3
- [3] Agata Brajdic and Robert Harle. Walk detection and step counting on unconstrained smartphones. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 225–234, 2013. 1, 2
- [4] Changhao Chen, Xiaoxuan Lu, Andrew Markham, and Niki Trigoni. Ionet: Learning to cure the curse of drift in inertial odometry. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2
- [5] Changhao Chen, Peijun Zhao, Chris Xiaoxuan Lu, Wei Wang, Andrew Markham, and Niki Trigoni. Oxiod: The dataset for deep inertial odometry. *arXiv preprint arXiv:1809.07491*, 2018. 2
- [6] Changan Chen, Kumar Ashutosh, Rohit Girdhar, David Harwath, and Kristen Grauman. Soundingactions: Learning how actions sound from narrated egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27252–27262, 2024. 3
- [7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, 2024. 4, 1
- [8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 2, 4
- [9] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. Scenefun3d: fine-grained functionality and affordance understanding in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14531–14542, 2024. 3
- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5
- [11] Zi-Yi Dou, Xitong Yang, Tushar Nagarajan, Huiyu Wang, Jing Huang, Nanyun Peng, Kris Kitani, and Fu-Jen Chu. Unlocking exocentric video-language data for egocentric video representation learning. *arXiv preprint arXiv:2408.03567*, 2024. 3
- [12] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 1, 2, 5
- [13] Hardegger et al. Actionslam: Using location-related actions as landmarks in pedestrian slam. In *IPIN*, 2012. 2
- [14] Zhang et al. Positioning method of pedestrian dead reckoning based on human activity recognition. In *IPIN*, 2022. 2
- [15] Chenyou Fan. Egovqa-an egocentric video question answering benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3
- [16] Xinyu Gong, Sreyas Mohan, Naina Dhingra, Jean-Charles Bazin, Yilei Li, Zhangyang Wang, and Rakesh Ranjan. Mmg-ego4d: Multimodal generalization in egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6481–6491, 2023. 2
- [17] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012, 2022. 2, 4
- [18] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19383–19400, 2024. 2, 4, 5
- [19] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE, 2024. 3
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5
- [21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 6
- [22] Sachini Herath, Hang Yan, and Yasutaka Furukawa. Ronin: Robust neural inertial navigation in the wild: Benchmark, evaluations, & new methods. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3146–3152. IEEE, 2020. 1, 2, 3, 5, 6, 8

- [23] Sachini Herath, Saghar Irandoust, Bowen Chen, Yiming Qian, Pyojin Kim, and Yasutaka Furukawa. Fusion-dhl: Wifi, imu, and floorplan fusion for dense history of locations in indoor environments. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5677–5683. IEEE, 2021. 2
- [24] Sachini Herath, David Caruso, Chen Liu, Yufan Chen, and Yasutaka Furukawa. Neural inertial localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6604–6613, 2022. 2, 3, 5, 6, 8
- [25] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22072–22086, 2024. 2
- [26] Yifei Huang, Jilan Xu, Baoqi Pei, Yuping He, Guo Chen, Lijin Yang, Xinyuan Chen, Yaohui Wang, Zheng Nie, Jinyao Liu, et al. Vinci: A real-time embodied smart assistant based on egocentric vision-language model. *arXiv preprint arXiv:2412.21080*, 2024. 3
- [27] Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video recap: Recursive captioning of hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18198–18208, 2024. 3
- [28] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2938–2946, 2015. 1
- [29] Kwan-Soo Kim and Yoan Shin. Deep learning-based pdr scheme that fuses smartphone sensors and gps location changes. *IEEE Access*, 9:158616–158631, 2021. 2
- [30] Mathieu Labbé and François Michaud. Rtab-map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation. *Journal of Field Robotics*, 36(2):416–446, 2019. 3
- [31] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visual-inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, 34(3):314–334, 2015. 2
- [32] Xin Li, Jian Wang, and Chunyan Liu. A bluetooth/pdr integration algorithm for an indoor positioning system. *Sensors*, 15(10):24862–24885, 2015. 2
- [33] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*, 2022. 3
- [34] Nie Lin, Takehiko Ohkawa, Yifei Huang, Mingfang Zhang, Minjie Cai, Ming Li, Ryosuke Furuta, and Yoichi Sato. Simhand: Mining similar hands for large-scale 3d hand pose pre-training. *arXiv preprint arXiv:2502.15251*, 2025. 2
- [35] Bonan Liu, Handi Yin, Manuel Kaufmann, Jinhao He, Sammy Christen, Jie Song, and Pan Hui. EgoHdm: An online egocentric-inertial human motion capture, localization, and dense mapping system. *arXiv preprint arXiv:2409.00343*, 2024. 2
- [36] Ruicong Liu, Takehiko Ohkawa, Mingfang Zhang, and Yoichi Sato. Single-to-dual-view adaptation for egocentric 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 677–686, 2024. 2
- [37] Wenxin Liu, David Caruso, Eddy Ilg, Jing Dong, Anastasios I Mourikis, Kostas Daniilidis, Vijay Kumar, and Jakob Engel. Tlio: Tight learned inertial odometry. *IEEE Robotics and Automation Letters*, 5(4):5653–5660, 2020. 1
- [38] Zihua Liu, Hiroki Sakuma, and Masatoshi Okutomi. VsrD: Instance-aware volumetric silhouette rendering for weakly supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17354–17363, 2024. 3
- [39] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004. 1
- [40] Shiyang Lu, Haonan Chang, Eric Pu Jing, Abdeslam Boularias, and Kostas Bekris. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In *Conference on Robot Learning*, pages 1610–1620. PMLR, 2023. 3
- [41] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy Bearman, and Babak Damavandi. Imu2clip: language-grounded motion sensor translation with multimodal contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13246–13253, 2023. 2, 6, 8
- [42] Anastasios I Mourikis and Stergios I Roumeliotis. A multi-state constraint kalman filter for vision-aided inertial navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3565–3572. IEEE, 2007. 2
- [43] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–544. Springer, 2022. 4, 5, 6, 7, 1
- [44] Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016. 6
- [45] Sebeom Park, Shokhrukh Bokijonov, and Yosoon Choi. Review of microsoft hololens applications over the past five years. *Applied sciences*, 11(16):7259, 2021. 1
- [46] Baoqi Pei, Yifei Huang, Jilan Xu, Guo Chen, Yuping He, Lijin Yang, Yali Wang, Weidi Xie, Yu Qiao, Fei Wu, and Limin Wang. Modeling fine-grained hand-object dynamics for egocentric video representation learning. In *International Conference on Learning Representations*, 2025. 3
- [47] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. EgoVlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5285–5297, 2023. 3

- [48] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30, 2017. 5
- [49] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. 2
- [50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PmlR, 2021. 3, 4, 6, 7, 1
- [51] Paul-Edouard Sarlin, Mihai Dusmanu, Johannes L Schönberger, Pablo Speciale, Lukas Gruber, Viktor Larsson, Ondrej Miksik, and Marc Pollefeys. Lamar: Benchmarking localization and mapping for augmented reality. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 686–704. Springer, 2022. 1, 2
- [52] Yuhan Shen, Huiyu Wang, Xitong Yang, Matt Feiszli, Ehsan Elhamifar, Lorenzo Torresani, and Effrosyni Mavroudi. Learning to segment referred objects from narrated egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14510–14520, 2024. 3
- [53] Yuanchao Shu, Kang G Shin, Tian He, and Jiming Chen. Last-mile navigation using smartphones. In *Proceedings of the 21st annual international conference on mobile computing and networking*, pages 512–524, 2015. 2
- [54] Scott Sun, Dennis Melamed, and Kris Kitani. Idol: Inertial deep orientation-estimation and localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6128–6137, 2021. 1
- [55] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023. 3
- [56] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 5, 6
- [57] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2820–2828, 2019. 5, 6
- [58] David Titterton and John L Weston. *Strapdown inertial navigation technology*. IET, 2004. 1
- [59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 5
- [60] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13525–13536, 2024. 3
- [61] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1189, 2023. 3
- [62] Hang Yan, Qi Shan, and Yasutaka Furukawa. Ridi: Robust imu double integration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 621–636, 2018. 2
- [63] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Vladislav Golyanik, Shaohua Pan, Christian Theobalt, and Feng Xu. EgoLocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors. *ACM Transactions on Graphics (TOG)*, 42(4):1–17, 2023. 2
- [64] Hong Ying, Carmen Silex, Andreas Schnitzer, Steffen Leonhardt, and Michael Schiek. Automatic step detection in the accelerometer signal. In *4th International Workshop on Wearable and Implantable Body Sensor Networks (BSN 2007) March 26–28, 2007 RWTH Aachen University, Germany*, pages 80–85. Springer, 2007. 1, 2
- [65] Ryo Yonetani, Jun Baba, and Yasutaka Furukawa. Retailopt: Opt-in, easy-to-deploy trajectory estimation from smartphone motion data and retail facility information. In *Proceedings of the 2024 ACM International Symposium on Wearable Computers*, pages 125–132, 2024. 2
- [66] Behnam Zeinali, Hadi Zanddizari, and Morris J Chang. Imunet: Efficient regression architecture for inertial imu navigation and positioning. *IEEE Transactions on Instrumentation and Measurement*, 2024. 1, 2, 5, 6
- [67] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Helping hands: An object-aware ego-centric video recognition model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13901–13912, 2023. 3
- [68] Mingfang Zhang, Jinglu Wang, Xiao Li, Yifei Huang, Yoichi Sato, and Yan Lu. Structural multiplane image: Bridging neural view synthesis and 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16707–16716, 2023. 3
- [69] Mingfang Zhang, Yifei Huang, Ruicong Liu, and Yoichi Sato. Masked video and body-worn imu autoencoder for egocentric action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 312–330. Springer, 2024. 2, 6
- [70] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6586–6597, 2023. 3